

# Fusing 2D Uncertainty and 3D Cues for Monocular Body Pose Estimation

Bugra Tekin      Pablo Márquez-Neila      Mathieu Salzmann      Pascal Fua  
EPFL, Switzerland

{bugra.tekin, pablo.marquezneila, mathieu.salzmann, pascal.fua}@epfl.ch

## Abstract

Most recent approaches to monocular 3D human pose estimation rely on Deep Learning. They typically involve training a network to regress from an image to either 3D joint coordinates directly, or 2D joint locations from which the 3D coordinates are inferred by a model-fitting procedure. The former takes advantage of 3D cues present in the images but rarely models uncertainty. By contrast, the latter often models 2D uncertainty, for example in the form of joint location heatmaps, but discards all the image information, such as texture, shading and depth cues, in the fitting step.

In this paper, we therefore propose to jointly model 2D uncertainty and leverage 3D image cues in a regression framework for monocular 3D human pose estimation. To this end, we introduce a novel two-stream deep architecture. One stream focuses on modeling uncertainty via probability maps of 2D joint locations and the other exploits 3D cues by directly acting on the image. We then study different approaches to fusing their outputs to obtain the final 3D prediction. Our experiments evidence in particular that our late-fusion mechanism improves upon the state-of-the-art by a large margin on standard 3D human pose estimation benchmarks.

## 1. Introduction

Monocular 3D human pose estimation is a longstanding Computer Vision problem. Over the years, two main classes of approaches have been proposed: Discriminative ones, that directly regress 3D pose from image data [2, 7, 32, 51, 62], and generative ones that search the pose space for a plausible skeleton configuration that aligns with the image data [20, 54, 63].

Recently, with the advent of ever larger datasets [28], models have evolved towards deep architectures, but the story remains largely unchanged. The state-of-the-art approaches can be roughly grouped into those that directly predict a 3D pose from images [28, 36, 59, 60], and those that first predict a 2D pose and then fit a 3D model to this 2D

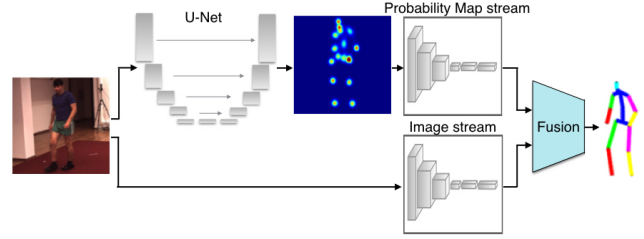


Figure 1. **Overview of our approach.** One stream of our network accounts for uncertainty by making use of probability maps of 2D joint locations. The second stream leverages all 3D cues in the input image by directly acting on it. The outputs of these two streams are then fused to obtain the final 3D human pose estimate. We study different fusion strategies ranging from early to late ones.

prediction [8, 67, 70]. While methods of the first kind leverage all the image information, such as texture, shading, and depth cues, they do not explicitly model body joint location uncertainty, which is critical to account for the ambiguities of 3D human pose estimation.

By contrast, methods of the second kind explicitly account for this uncertainty, for example by producing heatmaps for the expected 2D positions. However, they rarely use image information, such as depth cues, to guide the fitting process. The method of [37] is the only exception we know of. It searches for a 3D pose that best matches an embedding of the input image, previously learned with a Deep Network. In doing so, it does not rely on 2D pose, and can thus retain the relevant image cues. However, searching is done over the training data, which is slow and not particularly accurate.

In this paper, we introduce a discriminative approach that jointly leverages uncertainty, represented by 2D probability maps, along with all the cues present in the image, including 3D ones. To this end, we develop a two-stream Convolutional Neural Network (CNN) such as the one depicted by Fig. 1. Its first branch takes as input a probability map encoding the probable 2D joint locations and corresponding uncertainties. The probability map is itself computed using a U-shaped CNN [49] of the kind often used for semantic segmentation. The network’s second branch takes

the original image as input. The outputs of the two streams are combined by a fusion module that weighs their respective contributions and outputs a 3D pose. In short, our approach leverages the ability of one network to model 2D uncertainty and of the second to exploit 3D cues. Furthermore, it does not involve an expensive fitting procedure.

Ultimately, our key contribution is a general deep fusion framework to exploit both joint location uncertainty and 3D cues in the image. Here, we investigate several instances of this framework, corresponding to different fusion strategies ranging from early to late ones. To demonstrate the effectiveness of our approach, we evaluate these strategies on standard 3D human pose estimation benchmarks. Our experiments evidence the benefits of our approach over state-of-the-art methods, including both discriminative and generative ones. In particular, our late fusion strategy achieves significantly better accuracy than the state-of-the-art.

## 2. Related Work

Over the years, monocular 3D human pose estimation has received much attention in Computer Vision. The existing approaches can be roughly categorized into discriminative and generative ones. Here, we review both types of approaches, with a particular focus on the state-of-the-art.

Discriminative methods aim at predicting 3D pose directly from the input data, may it be single images [26, 27, 35, 36, 37, 47, 50, 59, 68], depth images [22, 45, 53], or short image sequences [60]. Early approaches falling into this category typically worked by extracting hand-crafted features and learning a mapping from these features to 3D poses [2, 7, 26, 27, 35, 51, 62]. Unsurprisingly, the more recent methods tend to rely on Deep Networks [36, 59, 60]. In particular, [36, 60] rely on 2D poses to pretrain the network, thus exploiting the commonalities between 2D and 3D pose estimation. In fact, [36] even proposes to jointly predict 2D and 3D poses. However, the two predictions are not coupled. More importantly, while these methods exploit all the available 3D image cues, they fail to model joint location uncertainty, which matters when addressing a problem as ambiguous as monocular 3D pose estimation.

Since pose estimation is much better-posed in 2D than in 3D, a popular way to model uncertainty is to use a generative model to find a 3D pose whose projection aligns with the 2D image data. In the past, this usually involved inferring a 3D human pose either by optimizing an energy function derived from image information, such as silhouettes [5, 13, 20, 21, 24, 29, 44, 54], feature trajectories [69] and 2D joint locations [3, 4, 19, 34, 46, 52, 63, 64], or 2D recognition-based pose retrieval approaches such as [17, 25, 39, 40]. In some algorithms [56, 57], the uncertainty was represented directly in the 3D pose space. With the growing availability of large datasets and the advent of Deep Learning, the emphasis has shifted towards using dis-

criminative 2D pose regressors [10, 12, 14, 15, 23, 30, 41, 42, 43, 61, 65, 66] to extract the 2D pose and infer a 3D one from it [8, 18, 67, 70]. The uncertainty is represented by heatmaps that encode the confidence of observing a particular joint at any given image location. A human body representation, such as a skeleton [70], or a more detailed model [8] can then be fitted to these predictions. While this takes uncertainty into account, it ignores image information during the fitting process. It therefore discards potentially important 3D cues that could help resolve ambiguities.

Among the methods that fit a 3D pose to the image data, the one of [37] is the only exception to this we know of. It relies on learning an image embedding whose inner product with the corresponding 3D pose is higher than with an unrelated one. The embedding does not rely on 2D pose, and can thus preserve 3D image cues. However, in the end, the 3D pose is obtained by searching over the training set for the pose that best matches the input image, which essentially amounts to a fitting procedure. This process is slow and relatively inaccurate, since it cannot generalize beyond the training data. Furthermore, while preserving image cues, the embedding does not explicitly model uncertainty.

Here, in contrast to earlier approaches, we propose to make the best of both worlds. We introduce a two-stream network that models both uncertainty, via a 2D probability map stream, and 3D image cues, via an image stream. Our experiments clearly demonstrate the importance of accounting for both these information sources.

## 3. Approach

Our goal is to increase the robustness and accuracy of 3D pose estimation from a single image by exploiting 3D image cues to the full while also accounting for joint location uncertainty. To this end, we rely on the two-stream architecture depicted by Fig. 1. The first stream operates on 2D probability maps that encode both the 2D joint locations and the corresponding uncertainties. The second extracts information directly from the original image. Their outputs are fused to predict a 3D pose. In the remainder of this section, we first formalize this general architecture. We then propose four different instances corresponding to different fusion strategies. Finally, we discuss how we compute the probability maps from the original image.

### 3.1. Formalization

Let  $\mathbf{I} : [1, W] \times [1, H] \times [1, 3] \rightarrow [0, 1]$  be the input RGB image,  $\mathbf{X} : [1, W] \times [1, H] \times [1, J] \rightarrow [0, 1]$  the probability maps encoding the probability of observing each one of  $J$  body joints at any given image location, and  $\mathbf{y} \in \mathbb{R}^{3J}$  the vector of 3D joint locations. Our two-stream architecture, as depicted by Fig. 1, comprises three main building blocks.

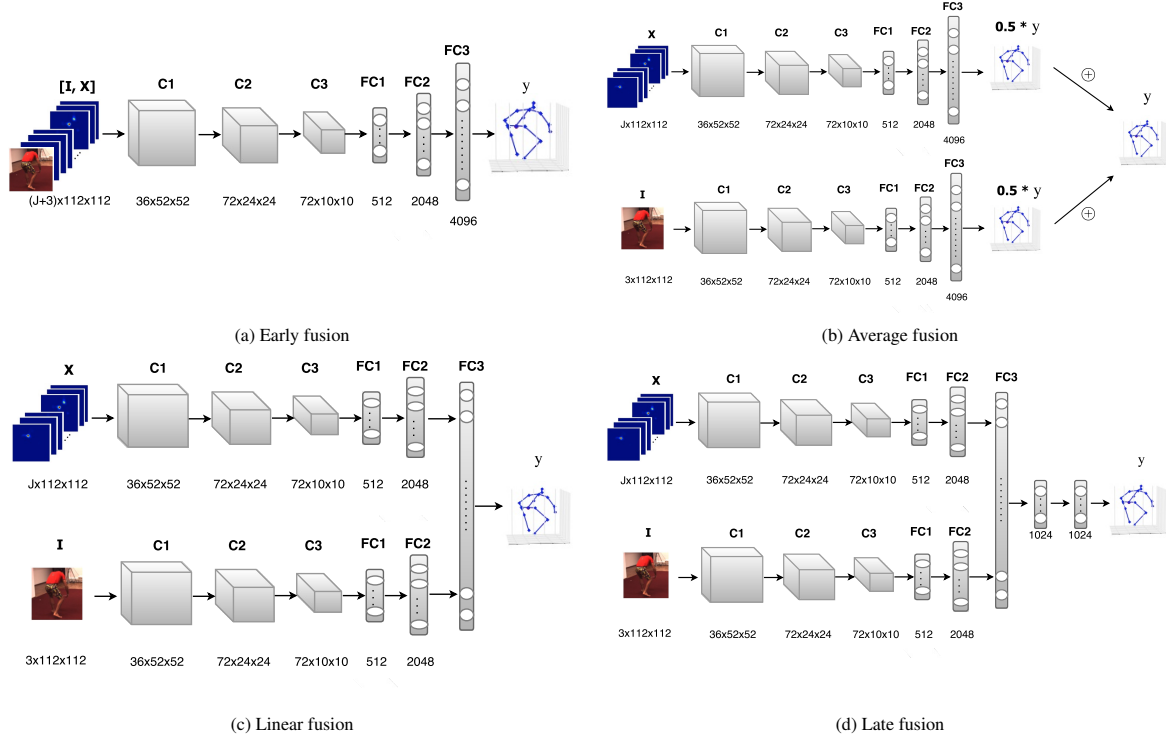


Figure 2. **Four different instances of our general fusion framework.** The four different fusion strategies depicted here all follow the pattern shown in Fig. 1. They combine 2D joint location probability maps with 3D cues directly extracted from the input image. In these four networks, we denote by  $C$  the convolutional layers and by  $FC$  the fully-connected ones. The numbers below each layer represent the corresponding size of the feature map for convolutional layers and the number of neurons for fully connected ones.

**Probability map stream.** It takes the probability map  $\mathbf{X}$  as input and returns

$$\mathbf{z}^X = h(\mathbf{X}; \theta_h), \quad (1)$$

where the behavior of function  $h(\cdot)$  is controlled by  $\theta_h$ . Here,  $\mathbf{z}^X$  denotes the output feature map. The probability map  $\mathbf{X}$  itself is estimated by a fully-convolutional network.

**Image stream.** It takes the image  $\mathbf{I}$  as input and returns

$$\mathbf{z}^I = g(\mathbf{I}; \theta_g), \quad (2)$$

where the behavior of function  $g(\cdot)$  is controlled by  $\theta_g$  and  $\mathbf{z}^I$  denotes the output feature map.

**Fusion Network.** It combines the outputs of  $h$  and  $g$  to predict the 3D pose. It can thus be expressed as

$$\hat{\mathbf{y}} = f(\mathbf{z}^X, \mathbf{z}^I; \theta_f), \quad (3)$$

where the behavior of function  $f(\cdot)$  is controlled by  $\theta_f$ .

Altogether, our two-stream network is therefore a composition of the functions  $h$ ,  $g$  and  $f$  that predicts a 3D pose from an image and corresponding probability maps given the parameters  $\theta = (\theta_h, \theta_g, \theta_f)$ . The output of this network can thus be written as

$$\hat{\mathbf{y}}(\mathbf{I}, \mathbf{X}; \theta) = f(h(\mathbf{X}; \theta_h), g(\mathbf{I}; \theta_g); \theta_f). \quad (4)$$

For training purposes, given a set of  $N$  training triplets  $(\mathbf{I}_i, \mathbf{X}_i, \mathbf{y}_i)$ , we learn the parameters  $\theta$  by minimizing the square loss, which can be expressed as

$$L(\theta) = \sum_{i=1}^N \|\hat{\mathbf{y}}(\mathbf{I}_i, \mathbf{X}_i; \theta) - \mathbf{y}_i\|_2^2. \quad (5)$$

We use the ADAM [33] gradient update method with a learning rate of  $10^{-3}$  to guide the optimization procedure. We rely on dropout with a probability of 0.5 after each fully-connected layer of the network and augment the training data by randomly cropping or rescaling  $112 \times 112$  patches from the  $128 \times 128$  input images to prevent overfitting and achieve translation invariance.

There are many ways to formulate the components  $h$ ,  $g$  and  $f$  either in terms of Deep Networks or of simple algebraic formulas so that the whole network can be trained end-to-end. We describe four of them below.

### 3.2. Two-Stream Architecture and Fusion

The goal of our two-stream network is to combine a notion of uncertainty on the 2D joint locations, coming from probability maps, with image cues providing information about 3D. Here, we propose four different strategies to fuse this information, which range from early to late fusion, and all fall into the formalism introduced above.

The four architectures corresponding to these strategies are shown in Fig. 2. They all use the same building blocks, that is, a CNN with three convolutional layers followed by three fully connected ones. The corresponding numbers of channels and feature map sizes are given in the figure. The filter sizes for the convolutional layers are  $9 \times 9$ ,  $5 \times 5$  and  $5 \times 5$ , respectively. We use a  $2 \times 2$  max-pooling layer after each convolutional layer. The activation function is the ReLU in all layers, except for the last one which has no nonlinearity.

The four architectures of Fig. 2 differ in the way the outputs of the two streams are fused. We describe these four approaches to fusion below.

**Early Fusion.** Since the image  $\mathbf{I}$  and the probability map  $\mathbf{X}$  have the same resolution  $W \times H$ , but different number of channels, the simplest approach to fusion is to concatenate them into a single  $W \times H \times (J + 3)$  volume that acts as input to a CNN trained to predict the 3D pose. In this case, the functions  $h(\cdot)$  and  $g(\cdot)$  of Eqs. 1 and 2 are simply identity maps. The fusion function  $f(\cdot)$  of Eq. 3 performs the concatenation and forward propagation through the CNN. Fig. 2(a) illustrates this strategy.

**Average Fusion.** At the other extreme, fusion can be performed by averaging 3D pose predictions from each stream. To this end, we implemented the model illustrated by Fig. 2(b). In this case,  $h(\cdot)$  represents the forward propagation of the probability map through a CNN that predicts the vector  $\mathbf{z}^X$  of Eq. 1, which here is a  $3J$ -dimensional vector representing a 3D pose. Similarly,  $g(\cdot)$  is implemented by another CNN that also predicts a  $3J$ -dimensional vector, but directly from the input image. The two CNNs have the same architecture but different weights. Fusion reduces to averaging these two pose estimates, that is,

$$f(\mathbf{z}^X, \mathbf{z}^I; \theta_f) = \frac{1}{2} (\mathbf{z}^X + \mathbf{z}^I). \quad (6)$$

Note that this is similar in spirit to the approach of [58] for action recognition, where the scores coming from an image stream and an optical-flow stream were averaged.

**Linear Fusion.** Average fusion, as described above, implicitly assumes the predictions from both streams to be equally reliable for all joints. In practice, there is no reason for this to be true. In fact, one expects the probability map stream, while possibly quite accurate in terms of 2D locations, to suffer from 3D ambiguities, which the image stream should help disambiguate, thanks to its access to subtle 3D image cues.

To account for this, we propose to weigh the respective contributions of each stream to the final pose prediction. To this end, we take the vectors  $\mathbf{z}^X$  and  $\mathbf{z}^I$  to be 4096-dimensional feature maps produced by the last fully-connected layers of the two streams that implement  $g$  and

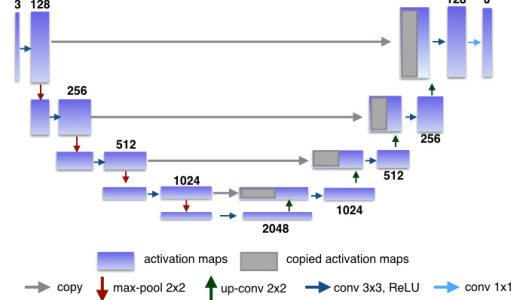


Figure 3. Network architecture for 2D probability map prediction. It consists of a contracting and expanding path where the features from the former are combined with the latter in order to exploit the context and localize joint positions. The network predicts probability maps, which encode the probability of a specific joint being observed at a given image location.

$h$ , and define  $f$  as

$$f(\mathbf{z}^X, \mathbf{z}^I; \theta_f) = \mathbf{W} \begin{bmatrix} \mathbf{z}^X \\ \mathbf{z}^I \end{bmatrix} + \mathbf{b}, \quad (7)$$

where the parameters  $\theta_f$  now include the  $3J \times 8192$  matrix  $\mathbf{W}$  and the bias  $\mathbf{b}$ . Fig. 2(c) illustrates this fusion strategy.

**Late Fusion.** Finally, we can go beyond the linear fusion of the vectors  $\mathbf{z}^X$  and  $\mathbf{z}^I$  described above and combine them in a nonlinear way. As shown in Fig. 2(d), we do this by introducing two additional fully-connected layers in the fusion function  $f(\cdot)$ . We use ReLU as the activation function to introduce nonlinearities. We will see in the results section that, in practice, this is the most effective approach.

### 3.3. 2D Joint Location Probability Map Prediction

Our approach depends on generating probabilistic maps of the 2D joint locations that we can feed as input to the probability map stream. To do so, we rely on a modified version of the U-Net of [49], which was initially developed for semantic segmentation in biomedical images and enables precise localization while capturing contextual information.

As shown in Fig. 3, it is a fully-convolutional architecture that includes links between non-consecutive layers. Given a  $W \times H \times 3$  RGB image  $\mathbf{I}$  as input, it performs a series of convolutions and pooling operations to reduce its spatial resolution, followed by upconvolutions to produce an output of the same resolution as the input image. In our case, this output is a  $W \times H$  probability map  $\mathbf{X}$  with  $J$  channels, each one of which encodes the probability of a specific joint to be observed at a given image location. In other words,  $\mathbf{X}(r, c, j)$  corresponds to the probability of finding the  $j^{\text{th}}$  joint at pixel  $(r, c)$ .

We modified slightly the original architecture [49]. First, for computational efficiency, we use a single convolution at every level, instead of two. Second, we doubled the number



of channels of the hidden feature maps to account for the larger number of channels of the output layer.

In its original formulation, the U-Net was designed to compute a separate probability distribution for every pixel over different channels, encoding the probability of a pixel to belong to one among several classes. Since we aim to compute a probability distribution per joint over the entire image, we modified the final softmax operation to reflect our goal. This yields a channel-wise softmax defined as

$$\mathbf{X}(r, c, j) = \frac{e^{\mathbf{L}(r, c, j)}}{\sum_{r', c'} e^{\mathbf{L}(r', c', j)}}, \quad (8)$$

where  $\mathbf{L}$  is the output of the last linear layer of the U-Net. This forces every channel of the output  $\mathbf{X}$  to be a probability distribution, meaning that  $\sum_{r, c} \mathbf{X}(r, c, j) = 1, \forall j$ .

During training, we leverage this property by using the average cross-entropy between every channel of  $\mathbf{X}$  and the ground-truth 2D positions  $\mathbf{y}^{2D}$  as our loss function. More specifically, given  $N$  training pairs  $(\mathbf{I}_i, \mathbf{y}_i^{2D})$ , the parameters of the network  $\theta_u$  are learned by minimizing the loss

$$L_u(\theta_u) = \frac{1}{NJ} \sum_{i=0}^N \sum_{j=0}^J H(\mathcal{N}(y_{ij}^{2D}, \sigma^2), \mathbf{X}_i(\cdot, \cdot, j)) \quad , \quad (9)$$

where we omitted the explicit dependency of  $\mathbf{X}$  on the parameters  $\theta_u$  to simplify the notation.  $\mathcal{N}(y_{ij}^{2D}, \sigma^2)$  is a normal distribution with mean  $y_{ij}^{2D}$  and variance  $\sigma^2$ , and  $H(p, q)$  is the standard cross-entropy given by

$$H(p, q) = - \sum_{r, c} p(r, c) \log q(r, c). \quad (10)$$

During training, we fix the standard deviation of the normal distribution to  $\sigma = 5$  pixels and use ADAM for parameter updates with a learning rate of  $10^{-3}$ .

In our experiments, we pretrained the U-Net for 2D probability map estimation as a preliminary step to training our two-stream network, using the training data specific to each experiment. Its parameters are then fixed, and we use it to generate the input to our two-stream network.

## 4. Results

In this section, we first describe the datasets we tested our approach on and the corresponding evaluation protocols. We then compare our approach against the state-of-the-art methods and provide a detailed analysis of our general framework.

### 4.1. Datasets

We evaluate our approach on the Human3.6m [28] and KTH Multiview Football II [9] datasets described below.

**Human3.6m** is a larger and more diverse motion capture dataset than its predecessors, such as HumanEva [55] and

the CMU Motion Capture Dataset [1]. It includes 3.6 million images with their corresponding 2D and 3D poses. The poses are viewed from 4 different camera angles. The subjects carry out complex motions corresponding to daily human activities. As in [36, 37, 70], we obtain the input images by extracting a square region around the subject using the bounding boxes that are part of the dataset and resize it to  $128 \times 128$ . We use the standard 17 joint skeleton from Human3.6m as our pose representation.

**KTH Multiview Football II** provides a benchmark to evaluate the performance of pose estimation algorithms in unconstrained outdoor settings. The camera follows a soccer player moving around the pitch. The videos are captured from 3 different camera viewpoints. As in Human3.6m, we resize the input images to  $128 \times 128$ . The output pose is a vector of 14 3D joint coordinates.

### 4.2. Evaluation Protocol

On Human3.6m, we used the same data partition as in earlier work [36, 37, 38, 60, 70] for a fair comparison. The data from 5 subjects (S1, S5, S6, S7, S8) was used for training and the data from 2 different subjects (S9, S11) was used for testing. We evaluate the accuracy of 3D human pose estimation in terms of average Euclidean distance between the predicted and ground-truth 3D joint positions, as in [36, 37, 38, 60, 70]. Training and testing were carried out monocularly in all camera views for each separate action.

In [8], the authors used a different protocol. Testing was carried out only in the frontal camera ("cam3") from trial 1 using the sequences from S9 and S11. The estimated skeleton was then further aligned to the ground-truth one by a rigid transformation. For completeness, we also evaluate our approach in this way.

On the KTH Multiview Football II dataset, we evaluate our method on the sequence containing Player 2, as in [6, 9, 60]. Following [60], the first half of the sequence from camera 1 is used for training and the second half for testing. To compare our results to those of [6, 9, 60], we report accuracy using the percentage of correctly estimated parts (PCP) score. Since the training set is quite small, we propose to pretrain our network on the recently released synthetic dataset [11], which contains images of sports players with their corresponding 3D poses. We then fine-tuned it using the actual training data from KTH Multiview Football II. We report results with and without this pretraining.

### 4.3. Comparison to the State-of-the-Art

We first compare our approach with state-of-the-art baselines on both datasets. Here, *Ours* refer to our late fusion strategy, which, as shown in Section 4.4, yields the best results among our four different strategies.

Input	Method	Directions	Discussion	Eating	Greeting	Phone Talk	Posing	Buying	Sitting	Sitting Down
Single-Image	Ionescu et al. [28]	132.71	183.55	132.37	164.39	162.12	150.61	171.31	151.57	243.03
	Li et al. [36]	-	148.79	104.01	127.17	-	-	-	-	-
	Li et al. [37]	-	134.13	97.37	122.33	-	-	-	-	-
	Li et al. [38]	-	133.51	97.60	120.41	-	-	-	-	-
	Zhou et al. [70]	-	-	-	-	-	-	-	-	-
	Rogez & Schmid [48]	-	-	-	-	-	-	-	-	-
	Tekin et al. [59]	-	129.06	91.43	121.68	-	-	-	-	-
Video	Tekin et al. [60]	102.41	147.72	88.83	125.28	118.02	112.3	129.17	138.89	224.90
	Zhou et al. [70]	87.36	109.31	87.05	103.16	<b>116.18</b>	106.88	99.78	124.52	199.23
	Du et al. [16]	85.07	112.68	104.90	122.05	139.08	105.93	166.16	117.49	226.94
Single-Image	Ours	<b>85.03</b>	<b>108.79</b>	<b>84.38</b>	<b>98.94</b>	<b>119.39</b>	<b>98.49</b>	<b>93.77</b>	<b>73.76</b>	<b>170.4</b>

Input	Method:	Smoking	Taking Photo	Waiting	Walking	Walking Dog	Walking Pair	Avg. (All)	Avg. (6 Actions)
Single-Image	Ionescu et al. [28]	162.14	205.94	170.69	96.60	177.13	127.88	162.14	159.99
	Li et al. [36]	-	189.08	-	77.60	146.59	-	-	132.20
	Li et al. [37]	-	166.15	-	68.51	132.51	-	-	120.17
	Li et al. [38]	-	163.33	-	73.66	135.15	-	-	121.55
	Zhou et al. [70]	-	-	-	-	-	-	120.99	-
	Rogez & Schmid [48]	-	-	-	-	-	-	121.20	-
	Tekin et al. [59]	-	162.17	-	65.75	130.53	-	-	116.77
Video	Tekin et al. [60]	118.42	182.73	138.75	<b>55.07</b>	126.29	<b>65.76</b>	124.97	120.99
	Zhou et al. [70]	107.42	143.32	118.09	79.39	114.23	97.70	113.01	106.07
	Du et al. [16]	120.02	135.91	117.65	99.26	137.36	106.54	126.47	118.69
Single-Image	Ours	<b>85.08</b>	<b>95.65</b>	<b>116.91</b>	<b>62.08</b>	<b>113.72</b>	<b>94.83</b>	<b>100.08</b>	<b>93.92</b>

Table 1. **Comparison of our approach with state-of-the-art algorithms on Human3.6m.** We report 3D joint position errors in mm, computed as the average Euclidean distance between the ground-truth and predicted joint positions. **Bold face** numbers denote the best overall methods, ***bold italic*** numbers denote the best methods among those only use single image as opposed to a sequence, if different. ‘-’ indicates that the results were not reported for the respective action class in the original paper. Note that our method consistently outperforms the baselines.

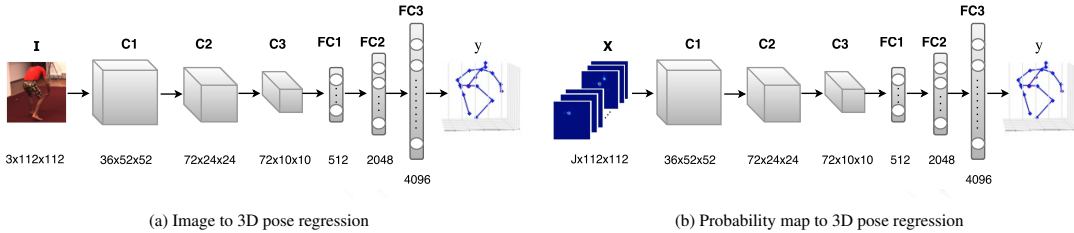


Figure 4. Baseline architectures we consider. **(a)** Regression from image to 3D human pose by a CNN, **(b)** Regression from probability maps to 3D human pose by a CNN. Its input is the joint location probability maps for 17 joints in the human body.

**Human3.6m.** In Table 1, we compare our results with those of the following state-of-the-art single-image based approaches: KDE regression from HOG features to 3D poses [28], jointly training a 2D body part detector and a 3D pose regressor [36], the maximum-margin structured learning framework of [37, 38], the deep structured prediction approach of [59] and 3D pose estimation with mocap guided data augmentation [48]. For completeness, we also compare our approach to the following methods that rely on either multiple consecutive images or impose temporal consistency: regression from short image sequences to 3D poses [60], fitting a sparse 3D pose model to 2D heatmap

predictions across frames [70], and fitting a 3D pose sequence to the 2D joints predicted by images and height-maps that encode the height of each pixel in the image with respect to a reference plane. [16].

As can be seen from the results in Table 1, our approach outperforms all the single-image baselines on all the action categories. In particular, it outperforms the image-based regression methods of [28, 36, 37, 38, 59], as well as the model-fitting strategy of [37, 38]. This, we believe, clearly evidences the benefits of fusing 2D uncertainty and 3D image cues, as achieved by our approach. Furthermore, we also achieve lower error than the method of [48], de-

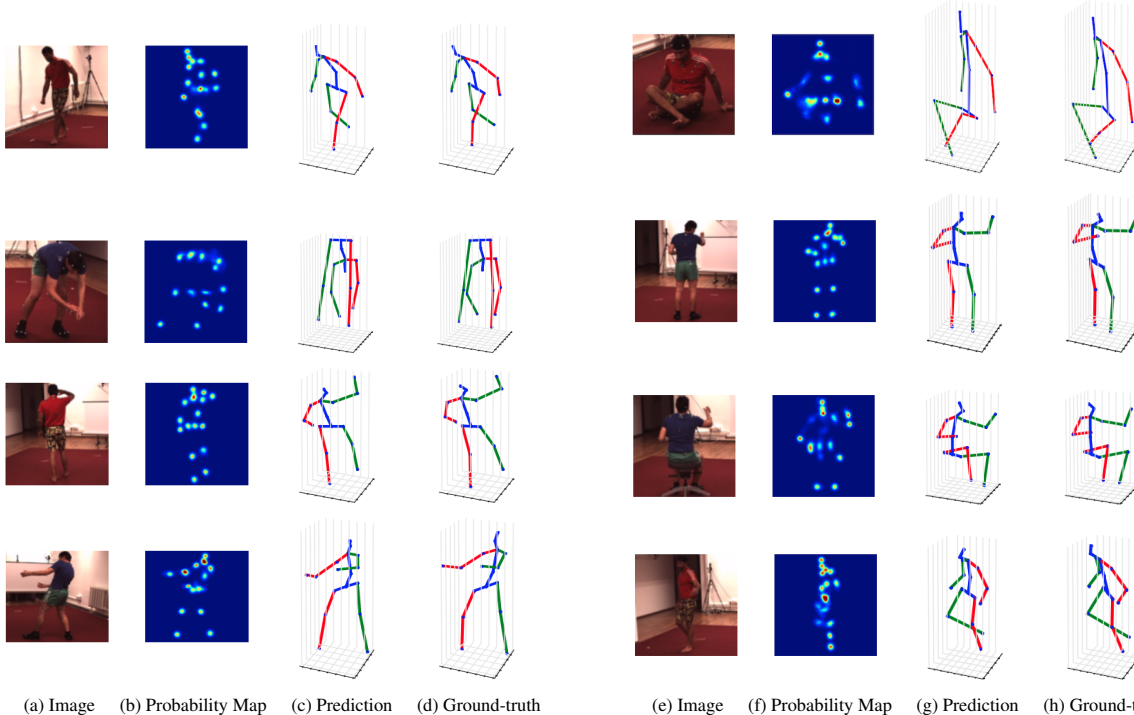


Figure 5. Pose estimation results on Human3.6m. **(a,e)** Input images. **(b,f)** 2D joint location probability maps. **(c,g)** Recovered pose. **(d,h)** Ground truth. Note that our method can recover the 3D pose in these challenging scenarios, which involve significant amounts of self occlusion and orientation ambiguity. Best viewed in color.

spite the fact that it relies on additional training data. Interestingly, even though our algorithm uses only individual images, it also outperforms the methods that rely on sequences [16, 60, 70] on most action categories. The fact that the methods of [60] and [70] are more accurate on a small subset of actions suggests that we could further improve our results by also enforcing consistency across frames, as they do. Fig. 5 depicts qualitatively some of our results.

As explained in Section 4.2, [8] reports pose estimation results only on the frontal camera. We carried out the same experiment and obtained an average 3D joint position error of 79.30 mm vs. 82.30 mm for [8]. Our approach therefore also outperforms [8], despite the fact that it fits a detailed 3D body model to 2D joint locations predicted with the state-of-the-art method of [43].

**KTH Multiview Football II.** In Table 2, we compare our approach on the KTH Multiview Football II dataset with the following state-of-the art methods: 3D pictorial structures [6, 9] and direct regression from image sequences [60]. Note that [6] and [9] rely on multiple views, and [60] makes use of video data. As discussed in Section 4.1, we report the results of two instances of our model: one trained on the standard KTH training data, and one pre-trained on the synthetic 3D human pose dataset of [11] and fine-tuned on the KTH dataset. Note that, while working

Method:	[9]	[9]	[6]	[60]	Ours-NoPretraining	Ours-Pretraining
Input:	Image	Image	Image	Video	Image	Image
Num. of cameras:	1	2	2	1	1	1
Pelvis	97	97	-	99	66	<b>100</b>
Torso	87	90	-	<b>100</b>	<b>100</b>	<b>100</b>
Upper arms	14	53	64	74	66.5	<b>100</b>
Lower arms	06	28	50	49	<b>100</b>	83
Upper legs	63	88	75	98	<b>100</b>	<b>100</b>
Lower legs	41	82	66	77	66.5	<b>83</b>
All parts	43	69	-	79	83.2	<b>93.2</b>

Table 2. On KTH Multiview Football II we compare our method that uses a single image to those of [9, 60] that use either one or two, the one of [6] that uses two, and the one of [60] that operates on a sequence. We rely on the percentage of correctly estimated parts (PCP) score to evaluate performance as in [6, 9, 60]. Higher PCP score corresponds to better 3D pose estimation accuracy.

with a single input image, both instances outperform all the baselines. Note also that pretraining on synthetic data yields the highest accuracy. We believe that this further demonstrates the generalization ability of our method. In Fig. 6, we provide a few representative poses predicted by our approach.

#### 4.4. Detailed Analysis

We now analyze two different aspects of our approach. First, we compare the different fusion strategies introduced in Section 3. In addition to these strategies, we also report

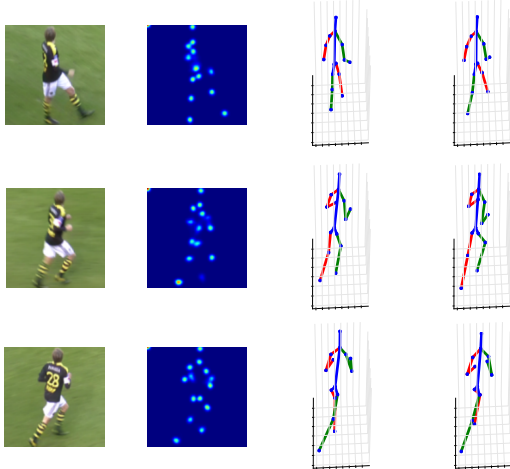


Figure 6. Pose estimation results on KTH Multiview Football II. In the first two columns, we show the input image and the predicted probability maps. First skeleton depicts our prediction and the second one depicts the ground-truth 3D pose. Best viewed in color.

the results of the following model-fitting baseline that enforces consistency of the projections of 3D poses and 2D joint uncertainties via an Expectation-Maximization (EM) framework similar to that of [70]: It consists of two different Deep Networks, one to predict 2D probability maps and one to predict 3D pose. The former is the same as our U-Net approach discussed in Section 3.3. The second one is a CNN with the same architecture as the image stream in Fig. 2(b), from which we estimate a density in 3D using Gaussian distributions around the predicted joint locations. Given these two predictions, we estimate the 3D pose by using an EM algorithm that couples 2D uncertainties and projection of 3D joint distributions. We will refer to this baseline as *EM-Optimization*.

The second aspect of our approach we analyze is the benefits of leveraging both 2D uncertainty and 3D cues. To this end, we make use of two additional baselines. The first one consists of a direct CNN regressor operating on the image only, as depicted in Fig. 4(a). We refer to this baseline as *Image-Only*. By contrast, the second baseline corresponds to a CNN trained to predict 3D pose from only the 2D probability maps (PM) obtained with our U-Net method, as shown in Fig. 4(b). We refer to this baseline as *PM-Only*.

In Table 3, we report the average pose estimation errors on Human3.6m for all these different methods. As mentioned before, our late fusion strategy yields the best results. Note, however, that all our fusion strategies outperform the state-of-the-art methods in Table 1. They also outperform the *EM-Optimization* baseline, thus demonstrating the advantage of our approach over model-fitting. Importantly, the *Image-Only* and *PM-Only* baselines perform worse than our approach, and all fusion-based methods. This evidences the importance of fusing 2D uncertainty and 3D cues for

Method:	3D Pose Error
Image-Only	128.47
PM-Only	120.07
EM-Optimization	116.95
Early Fusion	114.62
Average Fusion	112.07
Linear Fusion	109.02
Late Fusion	<b>100.08</b>

Table 3. 3D joint position errors (in mm) for the baselines and fusion strategies introduced in 3.2. The fusion networks perform better than those that use only the image or only the probability map as input. Late fusion achieves the best accuracy overall.

2D Prediction	3D Prediction	3D Error
Zhou et al. [70]	Zhou et al. [70]	133.91
Ours	Zhou et al. [70]	129.15
Ours	Ours	<b>116.96</b>

Table 4. Average Euclidean distance in millimeters with different 2D and 3D prediction methods. We evaluate the influence of 2D probability map prediction in the 3D pose accuracy by comparing the different stages of the method of [70] to those of our method. We evaluated on the first 1966 frames of the sequence corresponding to Subject 9 performing *Posing* action on camera 1 in trial 1 as was done in the online test code of [70].

monocular pose estimation.

During our experiments, we have observed that our U-Net-based 2D prediction network, depicted in Fig. 3, yields very accurate probability maps. Specifically, it achieves a localization error of 7.14 pixel on average over all actions, which outperforms the 10.85 error reported in [70]. To verify that our better 3D results are not only due to these better 2D results, we evaluated the method of [70] using our probability maps as input with their publicly available code. In Table 4, we compare these results with ours. Note that we still outperform this approach even when it relies on our 2D probability maps. This demonstrates that our better 3D predictions are truly the results of our fusion strategy.

## 5. Conclusion

In this paper, we have proposed to fuse 2D uncertainty and 3D image cues for monocular 3D human pose estimation. To this end, we have introduced a two-stream Deep Network that computes representations of 2D joint probability maps and RGB images, and fuses them to predict 3D pose. Our experiments have demonstrated that our late fusion strategy significantly outperforms the state-of-the-art methods on standard 3D human pose estimation benchmarks. Our framework is general, and can be extended to incorporate other modalities. In the future, we therefore intend to study the influence of part segmentations and optical flow on human pose estimation, along with temporal consistency when working with image sequences.





Figure 7. Pose estimation results on Human3.6m. (a,e) Input images. (b,f) 2D joint location probability maps. (c,g) Recovered pose. (d,h) Ground truth. Note that our method can recover the 3D pose in these challenging scenarios, which involve significant amounts of self occlusion and orientation ambiguity. Best viewed in color.

## A. Appendix

Below, we present quantitative results on the HumanEva-I dataset. Furthermore, in Figs. 7, 8 and 9, we provide additional qualitative results for the Human3.6m, HumanEva and KTH Multiview Football II datasets. Finally, in Fig. 10, we further demonstrate that our regressor trained on the recently released synthetic dataset of [11] generalizes well to real images obtained from the Leeds Sports Pose dataset [31]. Additional qualitative results can be found in the accompanying videos.

## B. Evaluation on the HumanEva Dataset

In Table 5, we present the performance of our late-fusion approach on the HumanEva-I dataset [55]. We adopted the evaluation protocol described in [8, 56, 70] for a fair comparison. We trained our regressors on the training sequences

of Subject 1, 2 and 3. The training and testing were carried out only on Camera 1. The 2D probability map estimation network and the late-fusion network are pretrained on Human3.6m and fine-tuned on HumanEva. As in [8, 56, 70], we measure 3D pose error as the average joint-to-joint distance after alignment by a rigid transformation.

Method	S1	S2	S3	Average
Simo-Serra et al. [56]	65.1	48.6	73.5	62.4
Bogo et al. [8]	73.3	59.0	99.4	77.2
Zhou et al. [70]	34.2	30.9	49.1	38.07
Ours	<b>31.04</b>	<b>14.80</b>	<b>38.16</b>	<b>28.00</b>

Table 5. Quantitative results of our late-fusion approach on Walking sequences of the HumanEva-I dataset [55]. S1, S2 and S3 correspond to Subject 1, 2, and 3, respectively. The accuracy is reported in terms of average Euclidean distance (in mm) between the predicted and ground-truth 3D joint positions.

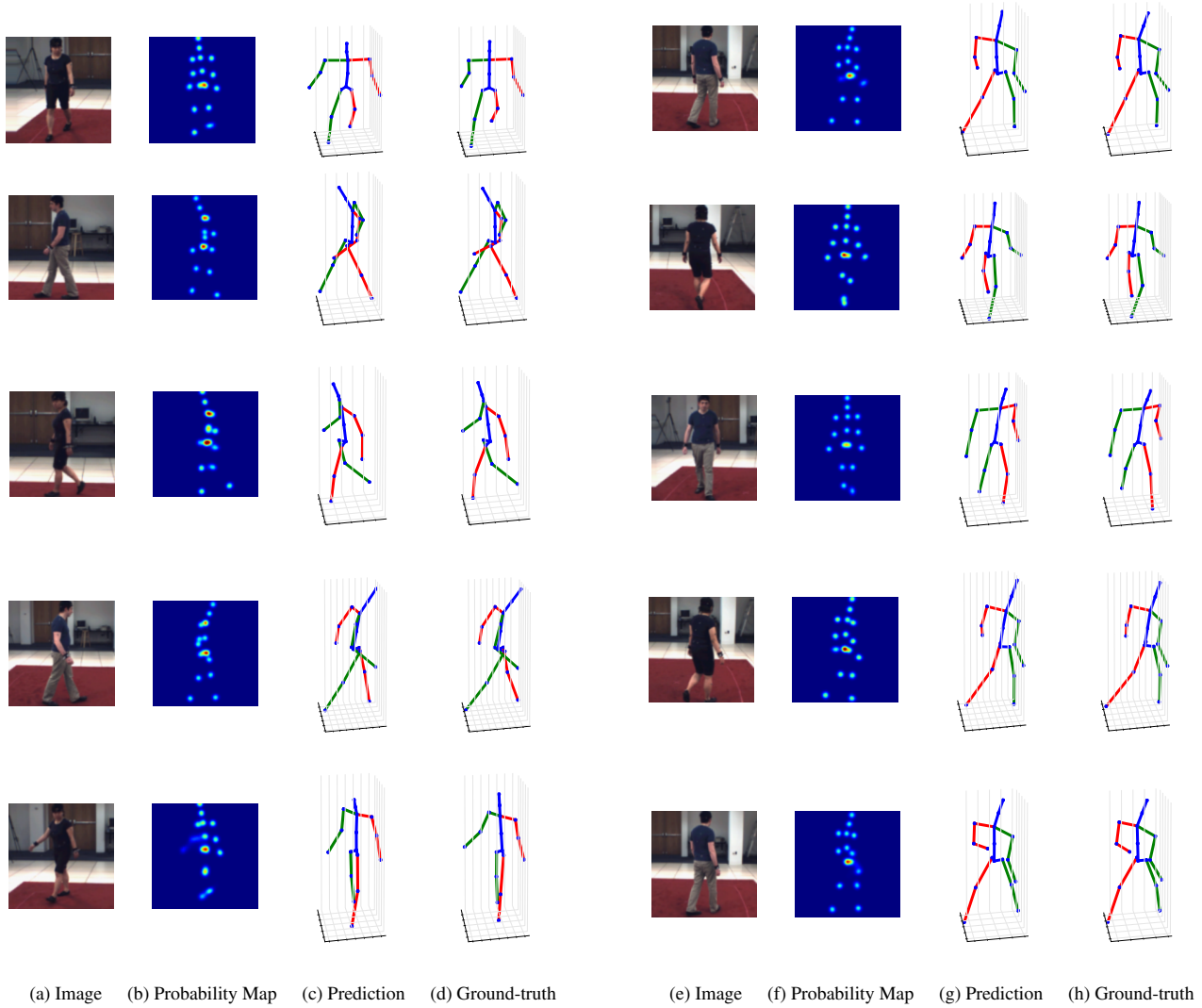


Figure 8. Pose estimation results on HumanEva-I. **(a,e)** Input images. **(b,f)** 2D joint location probability maps. **(c,g)** Recovered pose. **(d,h)** Ground truth. Best viewed in color.

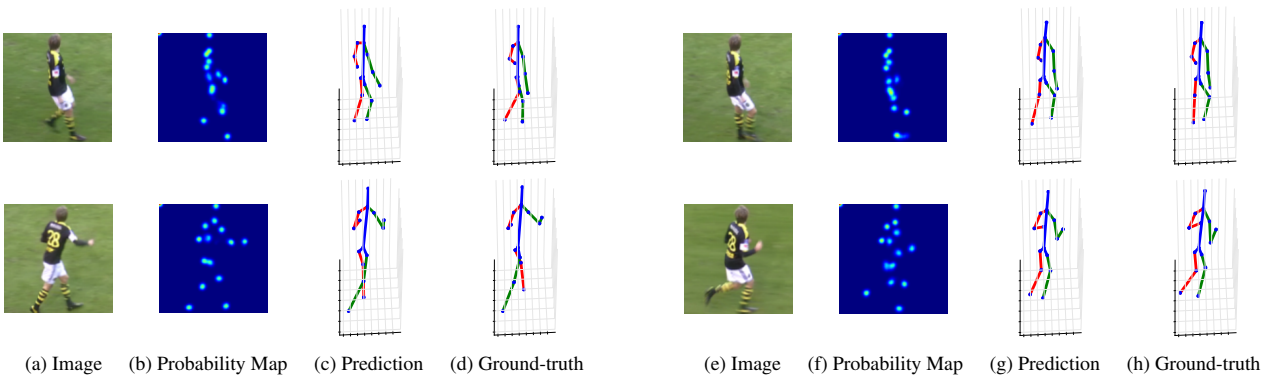


Figure 9. Pose estimation results on KTH Multiview Football II. **(a,e)** Input images. **(b,f)** 2D joint location probability maps. **(c,g)** Recovered pose. **(d,h)** Ground truth. Best viewed in color.

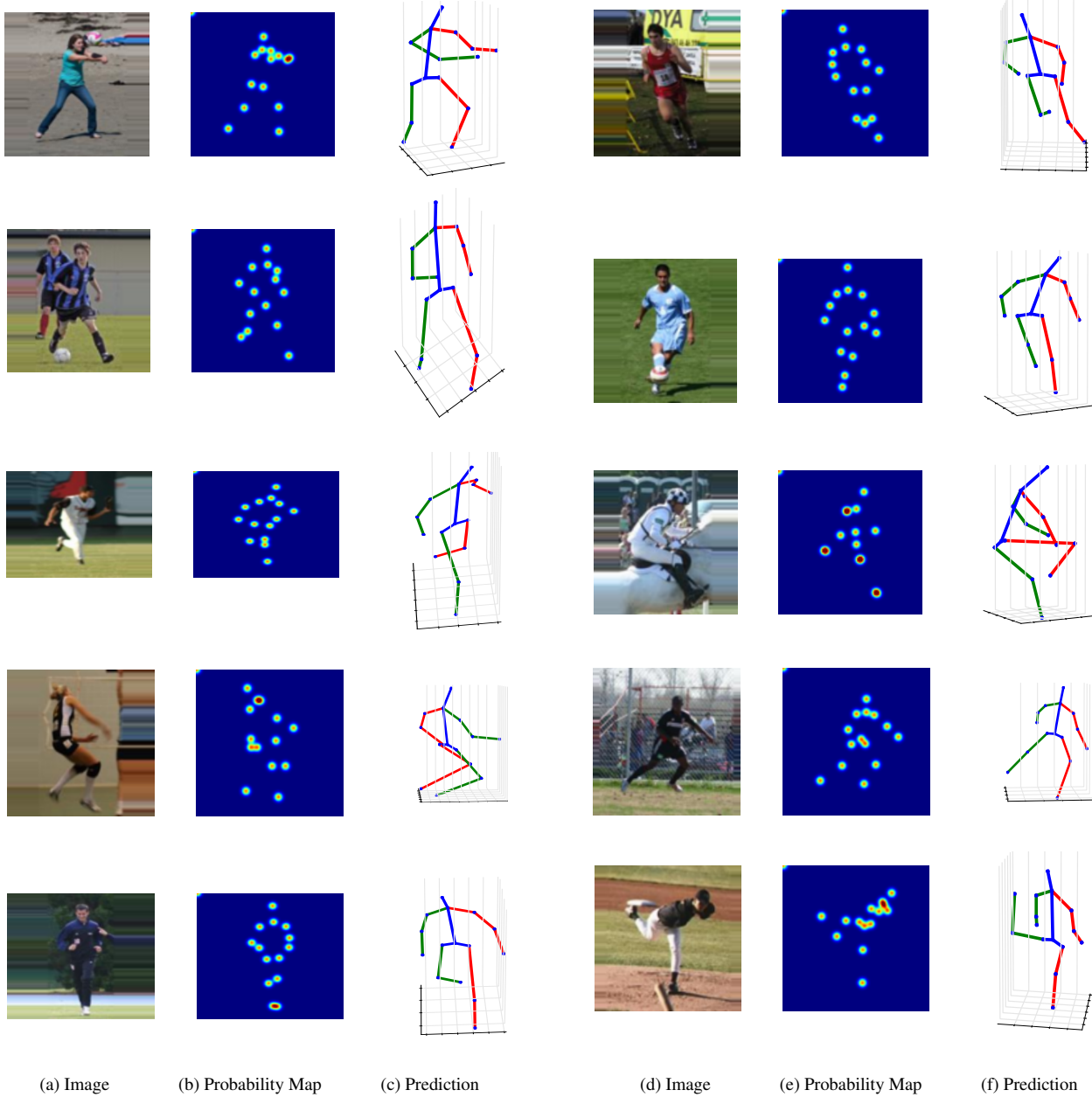


Figure 10. Pose estimation results on LSP. **(a,d)** Input images. **(b,e)** 2D joint location probability maps. **(c,f)** Recovered pose. We trained our network on the recently released synthetic dataset of [11] and tested it on the LSP dataset. The images were padded so as to be  $128 \times 128$  pixels. The quality of the 3D pose predictions demonstrates the generalization of our method. In the last row, we show failure cases due to background clutter and foreshortening. Best viewed in color.

## References

- [1] CMU Graphics Lab Motion Capture Database. <http://mocap.cs.cmu.edu/>. 5
- [2] A. Agarwal and B. Triggs. 3D Human Pose from Silhouettes by Relevance Vector Regression. In *CVPR*, 2004. 1, 2
- [3] I. Akhter and M. J. Black. Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction. In *CVPR*, 2015. 2
- [4] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D Pose Estimation and Tracking by Detection. In *CVPR*, 2010. 2
- [5] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed Human Shape and Pose from Images. In *CVPR*, 2007. 2
- [6] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D Pictorial Structures for Multiple Human Pose Estimation. In *CVPR*, 2014. 5, 7

- [7] L. Bo and C. Sminchisescu. Twin Gaussian Processes for Structured Prediction. *IJCV*, 2010. 1, 2
- [8] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 1, 2, 5, 7, 9
- [9] M. Burenius, J. Sullivan, and S. Carlsson. 3D Pictorial Structures for Multiple View Articulated Pose Estimation. In *CVPR*, 2013. 5, 7
- [10] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human Pose Estimation with Iterative Error Feedback. In *CVPR*, 2016. 2
- [11] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing Training Images for Boosting Human 3D Pose Estimation. In *3DV*, 2016. 5, 7, 9, 11
- [12] X. Chen and A. L. Yuille. Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. In *NIPS*, 2014. 2
- [13] Y. Chen, T. Kim, and R. Cipolla. Inferring 3D Shapes and Deformations from Single Views. In *ECCV*, 2010. 2
- [14] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured Feature Learning for Pose Estimation. In *CVPR*, 2016. 2
- [15] M. Du and R. Chellappa. Face Association Across Unconstrained Video Frames Using Conditional Random Fields. In *ECCV*, 2012. 2
- [16] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. Marker-less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps. In *ECCV*, 2016. 6, 7
- [17] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing Action at a Distance. In *ICCV*, pages 726–733, October 2003. 2
- [18] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient Convnet-Based Marker-Less Motion Capture in General Scenes with a Low Number of Cameras. In *CVPR*, 2015. 2
- [19] X. Fan, K. Zheng, Y. Zhou, and S. Wang. Pose Locality Constrained Representation for 3D Human Pose Reconstruction. In *ECCV*, 2014. 2
- [20] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and Filtering for Human Motion Capture. *IJCV*, 2010. 1, 2
- [21] S. Gammeter, A. Ess, T. Jaeggli, K. Schindler, B. Leibe, and L. Van Gool. Articulated Multi-Body Tracking Under Egomotion. In *ECCV*, 2008. 2
- [22] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient Regression of General-Activity Human Poses from Depth Images. In *ICCV*, 2011. 2
- [23] G. Gkioxari, A. Toshev, and N. Jaitly. Chained Predictions Using Convolutional Neural Networks. In *ECCV*, 2016. 2
- [24] P. Guan, A. Weiss, A. Balan, and M. Black. Estimating Human Shape and Pose from a Single Image. In *ICCV*, 2009. 2
- [25] N. R. Howe. A Recognition-Based Motion Capture Baseline on the Humaneva II Test Data. *MVA*, 2011. 2
- [26] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated Second-Order Label Sensitive Pooling for 3D Human Pose Estimation. In *CVPR*, 2014. 2
- [27] C. Ionescu, F. Li, and C. Sminchisescu. Latent Structured Models for Human Pose Estimation. In *ICCV*, 2011. 2
- [28] C. Ionescu, I. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 2014. 1, 5, 6
- [29] A. Jain, T. Thormahlen, H. Seidel, and C. Theobalt. MovieReshape: Tracking and Reshaping of Humans in Videos. In *SIGGRAPH*, 2010. 2
- [30] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler. Learning Human Pose Estimation Features with Convolutional Networks. In *ICLR*, 2014. 2
- [31] S. Johnson and M. Everingham. Clustered Pose and Non-linear Appearance Models for Human Pose Estimation. In *BMVC*, 2010. 9
- [32] A. Kanaujia, C. Sminchisescu, and D. N. Metaxas. Semi-Supervised Hierarchical Models for 3D Human Pose Reconstruction. In *CVPR*, 2007. 1
- [33] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimisation. In *ICLR*, 2015. 3
- [34] A. G. Kirk and J. F. O. D. A. Forsyth. Skeletal Parameter Estimation from Optical Motion Capture Data. In *CVPR*, 2005. 2
- [35] I. Kostrikov and J. Gall. Depth Sweep Regression Forests for Estimating 3D Human Pose from Images. In *BMVC*, 2014. 2
- [36] S. Li and A. Chan. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In *ACCV*, 2014. 1, 2, 5, 6
- [37] S. Li, W. Zhang, and A. B. Chan. Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation. In *ICCV*, 2015. 1, 2, 5, 6
- [38] S. Li, W. Zhang, and A. B. Chan. Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation. In *IJCV*, 2016. 5, 6
- [39] G. Mori and J. Malik. Estimating Human Body Configurations Using Shape Context Matching. In *ECCV*, 2002. 2
- [40] G. Mori and J. Malik. Recovering 3D Human Body Configurations Using Shape Contexts. *PAMI*, 2006. 2
- [41] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. In *ECCV*, 2016. 2
- [42] T. Pfister, J. Charles, and A. Zisserman. Flowing Convnets for Human Pose Estimation in Videos. In *ICCV*, 2015. 2
- [43] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *CVPR*, 2016. 2, 7
- [44] G. Pons-Moll, A. Baak, J. Gall, L. Leal-Taixe, M. Muller, H. Seidel, and B. Rosenhahn. Outdoor Human Motion Capture using Inverse Kinematics and von Mises-Fisher Sampling. In *ICCV*, 2011. 2
- [45] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon. Metric Regression Forests for Correspondence Estimation. *IJCV*, 2015. 2

- [46] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D Human Pose from 2D Image Landmarks. In *ECCV*, 2012. 2
- [47] G. Rogez, J. Rihan, C. Orrite, and P. Torr. Fast Human Pose Detection Using Randomized Hierarchical Cascades of Rejectors. 2012. 2
- [48] G. Rogez and C. Schmid. MoCap Guided Data Augmentation for 3D Pose Estimation in the Wild. In *NIPS*, 2016. 6
- [49] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 1, 4
- [50] R. Rosales and S. Sclaroff. Inferring Body Pose Without Tracking Body Parts. In *CVPR*, June 2000. 2
- [51] R. Rosales and S. Sclaroff. Learning Body Pose via Specialized Maps. In *NIPS*, 2002. 1, 2
- [52] M. Salzmann and R. Urtasun. Combining Discriminative and Generative Methods for 3D Deformable Surface and Articulated Pose Reconstruction. In *CVPR*, June 2010. 2
- [53] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient Human Pose Estimation from Single Depth Images. *PAMI*, 99, 2012. 2
- [54] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic Tracking of 3D Human Figures Using 2D Image Motion. In *ECCV*, 2000. 1, 2
- [55] L. Sigal and M. Black. Humaneva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion. Technical report, Department of Computer Science, Brown University, 2006. 5, 9
- [56] E. Simo-Serra, A. Quattoni, C. Torras, and F. moreno-noguer. A Joint Model for 2D and 3D Pose Estimation from a Single Image. In *CVPR*, 2013. 2, 9
- [57] E. Simo-Serra, A. Ramisa, G. Alenya, C. Torras, and F. moreno-noguer. Single Image 3D Human Pose Estimation from Noisy Observations. In *CVPR*, 2012. 2
- [58] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015. 4
- [59] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured Prediction of 3D Human Pose with Deep Neural Networks. In *BMVC*, 2016. 1, 2, 6
- [60] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In *CVPR*, 2016. 1, 2, 5, 6, 7
- [61] A. Toshev and C. Szegedy. Deeppose: Human Pose Estimation via Deep Neural Networks. In *CVPR*, 2014. 2
- [62] R. Urtasun and T. Darrell. Sparse Probabilistic Regression for Activity-Independent Human Pose Inference. In *CVPR*, 2008. 1, 2
- [63] R. Urtasun, D. Fleet, and P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *CVPR*, 2006. 1, 2
- [64] J. Valmadre and S. Lucey. Deterministic 3D Human Pose Estimation Using Rigid Structure. In *ECCV*, 2010. 2
- [65] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. In *CVPR*, 2016. 2
- [66] Y. Yang and D. Ramanan. Articulated Pose Estimation with Flexible Mixtures-Of-Parts. In *CVPR*, 2011. 2
- [67] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A Dual-Source Approach for 3D Pose Estimation from a Single Image. In *CVPR*, 2016. 1, 2
- [68] T.-H. Yu, T.-K. Kim, and R. Cipolla. Unconstrained Monocular 3D Human Pose Estimation by Action Detection and Cross Modality Regression Forest. In *CVPR*, 2013. 2
- [69] F. Zhou and F. de la Torre. Spatio-Temporal Matching for Human Detection in Video. In *ECCV*, 2014. 2
- [70] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *CVPR*, 2016. 1, 2, 5, 6, 7, 8, 9